A machine learning approach for Colles' fracture treatment diagnosis

Kwun Ho Ngan[1 [0000−0001−7623−942X]], Artur d'Avila Garcez[1 [0000−0001−7375−9518]], Karen M. Knapp[2 [0000−0002−7360−0926]], Andy Appelboam[3 [0000−0002−2982−9707]], and Constantino Carlos Reyes-Aldasoro[1 [0000−0002−9466−2018]]

[1] School of Mathematics, Computer Science and Engineering
City, University of London, London, EC1V 0HB, UK
{kwun-ho.ngan, a.garcez, reyes}@city.ac.uk
[2] College of Engineering, Mathematics and Physical Science
University of Exeter, Exeter, EX4 4QJ, UK k.m.knapp@exeter.ac.uk
[3] Royal Devon and Exeter Hospital Barrack Road, Exeter, EX2 5DW, UK
andy.appelboam@nhs.net

Abstract. Wrist fractures (e.g. Colles' fracture) are the most common injuries in the upper extremity treated in Emergency Departments. Treatment for most patients is an intervention called *Manipulation under Anaesthesia* (MUA). Surgical treatment would be needed for complex fractures or if the wrist stability is not restored. In addition, an unsuccessful treatment via MUA may also require subsequent surgical operation causing inefficiency in constrained medical resources and patients' inconvenience. Previous geometric measurements in X-ray images [21] were found to provide statistical differences between healthy controls and patients with fractures, as well as pre- and post-intervention images. The most discriminating measurements were associated with the texture analysis of the radial bone. This work presents further analysis of these measurements and applying them as features to identify an appropriate machine learning model for Colles' fracture treatment diagnosis. Random forest was evaluated to be the best model based on classification accuracy among the selected models commonly used in similar research. The non-linearity of the measurement features has attributed to the superior performance of an ensembled tree-based model. It is also interesting that the most important features (i.e. texture and swelling) required in the optimised random forest model are consistent with previous findings [21].

Keywords: Fracture · Radiography · Feature Analysis · Random Forest

1 Introduction

Colles' fracture is one of the most common fractures of the radial bone leading to posterior displacement of distal fragments at the wrist [8]. The fracture generally causes acute pain and swelling. It may even lead to residual impairment in hand and wrist motion if left untreated [13, 23].

Current evaluation of fracture severity and stability is primarily based on radiographs before and, when needed, after intervention [13]. A typical radiographic study will include analysis from multiple views (i.e. posteroanterior and lateral views) of the fractured wrist. Oblique views might also be needed to better define the fracture location.

The type of treatment generally depends on fracture displacement, angulation and shortening [13]. Fracture treatment has evolved over the past two decades improving stability and anatomical functionality. The advancement has also reduced the risk of complications such as neuropathies, arthrosis, tendon ruptures and finger stiffness during the rehabilitation following a distal radius fracture [9, 23]. Minor fractures (such as extra-articular, stable and minimally displacement with no comminution) are generally treated at the Emergency Department using closed reduction and immobilization technique (e.g. manipulation under anaesthesia (MUA) [2, 16]. An unsuccessful treatment of MUA would however require subsequent surgical operation (e.g. open reduction and internal fixation (ORIF)).

Despite considerable research on the appropriate intervention according to fracture characteristics [1–3, 11, 15], the choice of treatment remains highly subjective to the X-ray interpretation by the radiologist and/or surgeon, and largely depends upon the available clinical information on a case-by-case basis.

This work aims to develop a classification model based on a dataset of X-ray images taken in an Emergency Department to detect fractured wrist cases from the healthy controls. Within the fractured cases, it will further aim to determine the success of closed reduction treatment (i.e. MUA). Effect of feature groups are also evaluated to verify the findings from previous study [21] and utilise these findings for more precise diagnostics models in the future.

2 Materials and Methods

2.1 Patient Dataset

A data collection of 161 independent cases of wrist fracture was used in this work. The data was sourced ethically from the Royal Devon and Exeter Hospital with Caldicott Guardian approval. Prior informed consent was obtained from each participant. The data was subsequently anonymised according to the ethics policy of the donating institution.

Each case included basic anonymised patient information (e.g. age and gender), as well as, human-annotated measurements from X-ray radiographs in posterior-anterior (PA) views. The selection of these measurements has been documented in [21]. There were 139 cases diagnosed with wrist fracture of varying severity and 22 cases of healthy controls. Among the cases of fracture, it was also retrospectively classified into sub-groups of pre-successful (n=50), pre-unsuccessful (n=31), post-successful (n=40) and post-unsuccessful (n=18) based on patient electronic records. *Pre-* and *post-* refer to before and after the intervention. *Successful* refers to those cases for which the intervention (MUA) was

successful and no further treatment was required, and *unsuccessful* for those cases where surgery was recommended, but not yet performed.

2.2 Acquisition of X-ray Radiographs

The X-ray images were obtained with the following systems across a range of exposure factors:

1. DigitalDiagnost DidiEleva01 (Philips Medical Systems, Netherlands)
2. Mobile tablet work station (Thales, France)
3. DirectView CR 975 (Kodak, USA)
4. DirectView CD850A (Kodak, USA)
5. Definium 5000 (GE Healthcare, USA)

The raw images were stored in DICOM format [4]. Some representative cases of these images are shown in Fig. 1. These images are expected to represent the variation in image quality, positioning of wrist and presence of noisy input (e.g. collimation lines and text legends).

Fig. 1. Six representative X-ray images were collected from previous clinical activity at Royal Devon and Exeter NHS Foundation Trust Emergency Department. These images present the variability in the quality, positioning of the arm and presence of noisy input (e.g. collimation lines and text legends). The images were anonymised while metadata such as age, date of acquisition, gender and clinical outcome were retained.

2.3 Image Pre-processing and Model Feature Generation

Three landmark points were firstly located manually - (1) base of the lunate, (2) extreme of the radial styloid and (3) centre of the metacarpal of the middle finger.

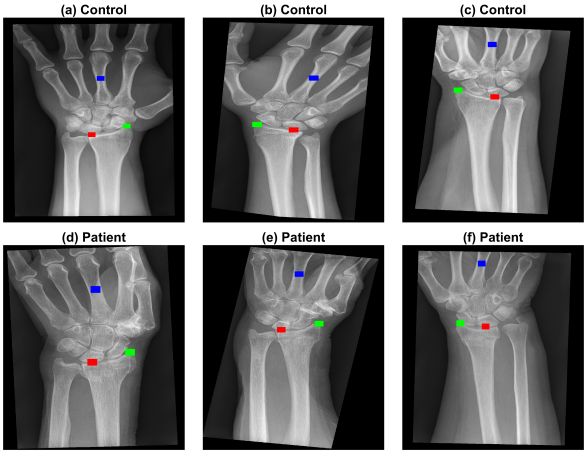

Fig. 2. Automatic pre-processing of the radiographs and location of landmarks. The six representative cases shown in Fig. 1 were automatically rotated so that the forearm is vertical, the artefacts due to the collimator were removed and landmarks for lunate, radial styloid and metacarpal were manually located.

The X-ray images were pre-processed automatically using Matlab based on the identified landmarks. A detailed description of the pre-processing procedure can be referred to [21]. This procedure is summarised below.

Each image was firstly aligned vertically along the forearm. The lines of collimation were also removed. Fig. 2 shows some of the pre-processed images. The following three indicator groups of measurements were extracted from the processed images to determine the appropriate treatment procedures. An illustration of geometric measurements from [21] is presented in Fig. 3.

1. **Indication of Swelling:** Widths along the forearm as an indication of swelling.
2. **Indication of Osteoporosis:** Width of finger and ratio of trabecular area to the total area of interest measured at the middle finger [6, 24, 14].
3. **Texture Indication at Radial Bone:** Texture measurement from a selected region of radial bone using Local Binary Pattern (LBPs) [19]. Grayscale intensity profile across landmarks #2 and #3 to indicate the changes in texture in the radial bone.

This forms a feature set of 36 variables for model construction including age, gender and 34 image-based measurements in the three indicator groups. Table 1 presents the exact features used in this study.

2.4 Predictive Classification Models

Three different classification models were used for the selection of the best predictive model - Logistic Regression (LR), Decision Tree (DT) and Random Forest

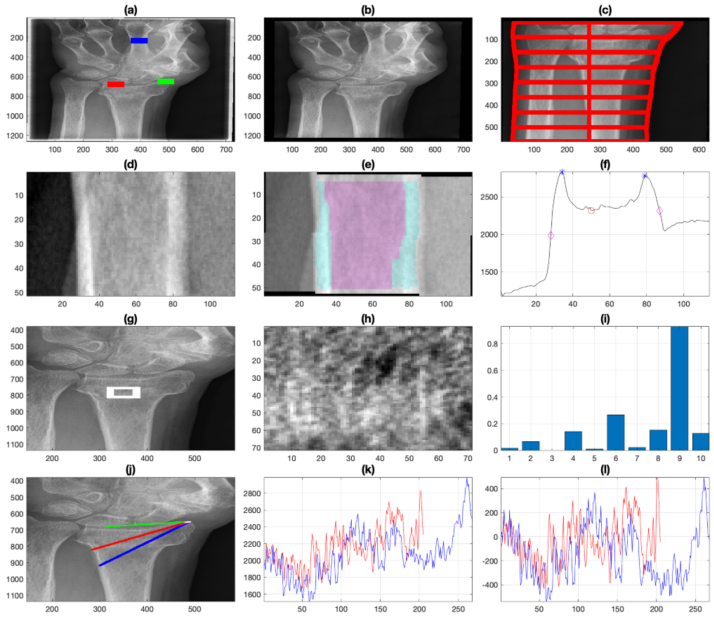

Fig. 3. Illustration of preliminary characterisation of radiographs with geometric measurements. (a) Original radiograph with landmarks. (b) Pre-processing. (c) Boundaries and traces of forearm. (d) Cortical bone region of interest (ROI). (e) Segmentation of cortical and trabecular bone regions. (f) Intensity profile of the average intensity projection of the bone with the following markers: edges of the bone (magenta diamond), limit of cortical bone (blue square) and centre of bone (red circle). (g) Texture ROI on the radius. (h) Zoom into the texture ROI. (i) Ten texture coefficients generated by Local Binary Pattern analysis. (j) Two profile lines from the radial styloid: at 30 (red) and at 45 (blue) degrees from the line (green) between the lunate the radial styloid landmarks. (k) Intensity profiles corresponding to the lines traced in (j). (l) Intensity profiles adjusted by removing the slope. [21]

(RF). These models have been widely used in classification for medical use cases [5, 7, 10, 18] and have found varying success in prediction.

Logistic Regression is a linear model for classification based on the probabilities describing the possibility of outcomes (i.e. classes) of a single sample using a logistic function [5, 20]. In a multi-class scenario, the class with the highest probability denotes the predicted class [22]. Decision Trees are non-parametric classification models by learning simple decision rules from features to predict class value [20]. Random Forest is an ensemble classifier by fitting a pre-defined number of decision trees and finding the most probable class from the average of the leaf nodes of the decision trees [17, 20].

Given that the number of samples (n=161) in the dataset is limited, oversampling has been applied in this study to mimic sampling from the population by

Grouping of Features for Treatment Classification			
Basic Patient Info	**Indication of Swelling**	**Indication of Osteoporosis**	**Texture Indication at Radial Bone**
Age	Ratio: Width Line 1 / Width Line 4 (WW1)	Ratio: Trabecular Area / Total Area (TrabecularToTotal)	Local Binary Pattern 1 (LBP1)
Gender	Ratio: Width Line 2 / Width Line 4 (WW2)	Finger Width (WidthFinger)	Local Binary Pattern 2 (LBP2)
	Ratio: Width Line 3 / Width Line 4 (WW3)		Local Binary Pattern 3 (LBP3)
	Ratio: Width Line 5 / Width Line 4 (WW5)		Local Binary Pattern 4 (LBP4)
	Ratio: Width Line 6 / Width Line 4 (WW6)		Local Binary Pattern 5 (LBP5)
	Ratio: Width Line 7 / Width Line 4 (WW7)		Local Binary Pattern 6 (LBP6)
	Ratio: Width Line 8 / Width Line 4 (WW8)		Local Binary Pattern 7 (LBP7)
	Ratio: Min Width Line / Max Width Line (MinMaxWidthAtCM)		Local Binary Pattern 8 (LBP8)
	Ratio: Width Line 1 + 8 / Width Lines 4 + 5 (W1W8/W4W5)		Local Binary Pattern 9 (LBP9)
	Ratio: Width Line 1 + 2 / Width Lines 7 + 8 (W1W2/W7W8)		Local Binary Pattern 10 (LBP10)
			Average Row LBP (stats.row_LBP)
			Average Column LBP (stats.col_LBP)
			Slope First Profile Full Line (stats.slope_1)
			Slope Second Profile Full Line (stats.slope_2)
			Slope First Profile Short Segment (stats. slope_short_1)
			Slope Second Profile Short Segment (stats. slope_short_2)
			Standard Deviation First Profile (stats.std_1)
			Standard Deviation Second Profile (stats. std_2)
			Standard Deviation First Profile Adjusted (stats.std_ad_1)
			Standard Deviation Second Profile Adjusted (stats.std_ad_2)
			Distance First Profile (dist1)
			Distance Second Profile (dist2)

Table 1: Grouping of Features for Treatment classification by Indicator Type. Details of measurements for each individual feature can be found in [21].

data replication using sampling with replacement. The resulting sample dataset had a size of 10000 samples where further increase in sample sizing did not appear to yield significant model improvement. The distribution of the 5 classes in the oversampled dataset remained consistent. This was based on the assumption that the dataset of X-ray images was representative of the distribution of batch analysed images within a typical Emergency Department and has a natural bias towards fractured cases.

This work was written in Python and used the Scikit-Learn package for model construction and evaluation. Each model was trained and evaluated in a 10-fold cross-validation. Default values for model hyperparameters can be assumed unless otherwise stated in the model evaluation. The average classification accuracy of each model was used for performance comparison. All models were set to the same random state to ensure comparability.

3 Analysis and Evaluation

3.1 Exploratory Data Analysis

The cross-correlation between the given features was first investigated. A heatmap (Fig. 4) was generated to determine the degree of correlation. It can be observed that there is a strong correlation between features within indicator groups (e.g.

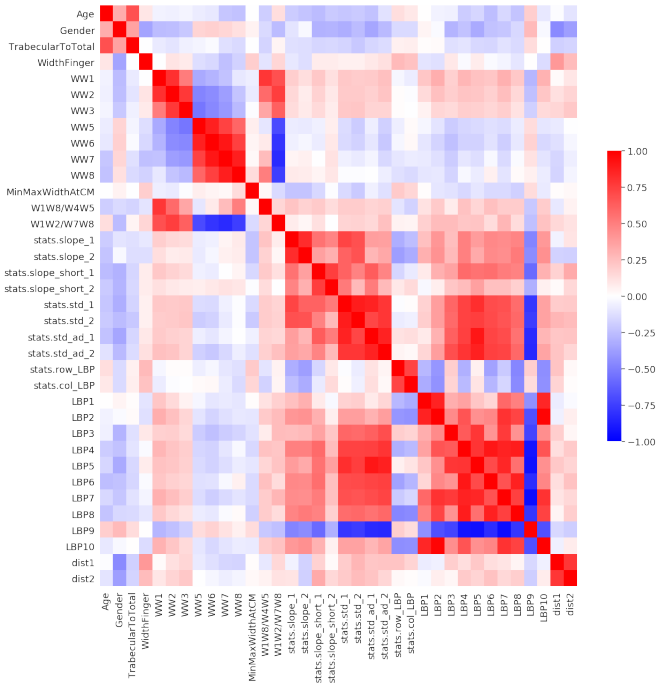

Fig. 4. Heatmap showing clusters of cross-correlations among variables in the Fracture dataset. Blue indicates a negative correlation and red indicates a positive correlation. The colour intensity of each cell shows the strength of the correlation.

wrist width ratios as swelling indicators, local binary profiles and grayscale intensity profiles on the texture of the radial bones).

Interestingly, LBP9 has a distinctive negative correlation with other texture related features. Given the bin value range for LBP9 (approx. 204 to 229 in binary format), it may indicate the contrasting presence of edges (e.g. hairline cracks) and smooth surface (e.g. blocks of whiteness of a cast).

In addition, the low cross-correlation across different indicator groups suggests that these groups were relatively independent. This was useful to assess the significance of the indicator groups on classification performance that will be discussed later in this work.

To understand the univariate correlation for the classification task, distribution plots of all the features (except LBP3 & LBP5) are presented in Fig. 5. LBP3 and LBP5 were omitted due to the predominant values of zeros (0) across the cases causing a distribution calculation error. It can be observed that some features (e.g. Width of Finger) showed no significant difference across classes while others (e.g. Local Binary Pattern) showed different distribution between pre-treatment and post-treatment conditions. Age and Radial Bone Landmark Distance (dist_2) also showed differences between healthy wrist and fractured

wrist. No feature showed a significant difference in distribution between successful and unsuccessful treatments. These findings were consistent with the findings published in [21].

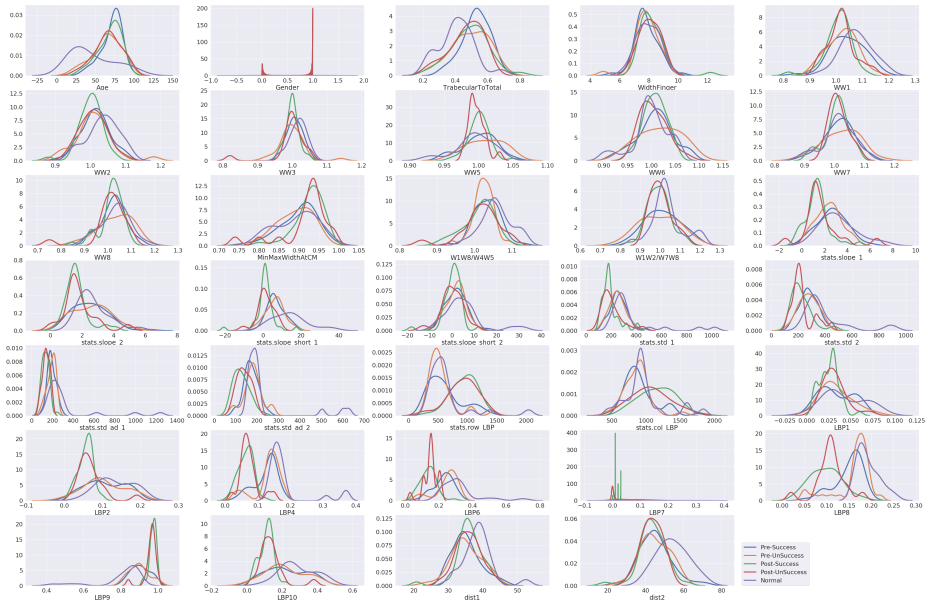

Fig. 5. Univariate distribution by treatment classes to identify prominent features for classification.

3.2 Model Evaluation

Based on the initial exploratory data analysis, univariate correlation for classification had been poor. This had implication to the model performance accuracy used in this work. Fig. 6 shows the summary results of the 3 models over 10-fold cross-validation. The poor univariate correlations and linear separability could be attributing to the poor performance for the logistic regression model (Ave. Acc.: 71.0% & weighted F1-Score: 0.69).

The accuracy of a decision tree showed significant improvement (Ave. Acc.: 86.7% & weighted F1 Score: 0.87). The tree branches allowed better classification through non-linear separation. A maximum tree depth of 6 (i.e. square root of the total number of features) was used as recommended in [12] to avoid overfitting. Fig. 7 presents a sample distribution plot of a leaf node in a decision tree where post-treatment samples were separated entirely, and a threshold can be seen for separating the healthy cases from the fractured pre-treatment samples.

To further boost the accuracy of the model, a random forest of 4 decision trees was selected with the same tree hyper-parameters as the decision tree model to

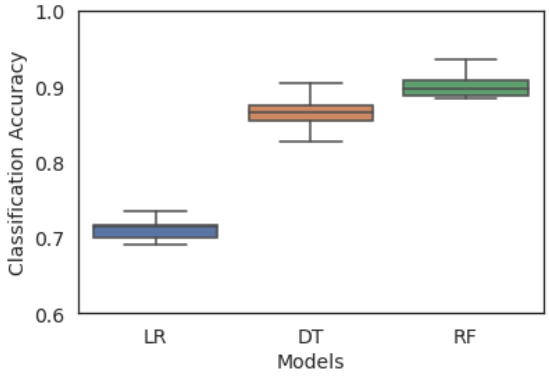

Fig. 6. Classification Performance for Logistic Regression (LR), Decision Tree (DT) and Random Forest (RF) models over 10-fold cross-validation based on classification accuracy of prediction.

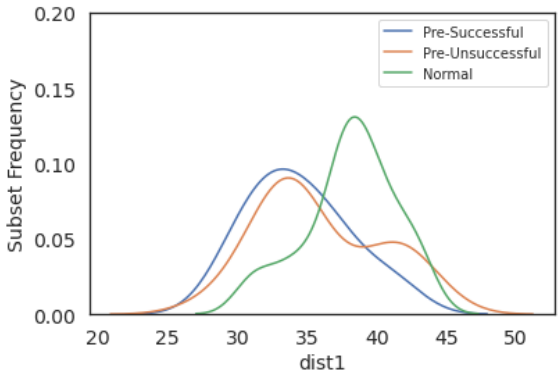

Fig. 7. Distribution of radial bone landmark distance (dist_1) within a leaf node (Local Binary Pattern 4 > 0.09 & Local Binary Pattern 8 > 0.16 & Ratio of Wrist Line 2 over Wrist Line 4 <= 1.04).

demonstrate a significant improvement in classification accuracy of more than 3% to an accuracy of 90.0% (weighted F1 Score: 0.94). Increasing the number of trees will potentially improve the accuracy further with added model complexity. This is not necessary for the illustration in this study. The aggregation of weak classifications from each tree enhanced the probability of the most expected class.

The impact of different indicator groups on classification accuracy was also investigated. Fig. 8 presents the accuracy of the same configured random forest model trained using different subsets of features. It can be clearly observed that

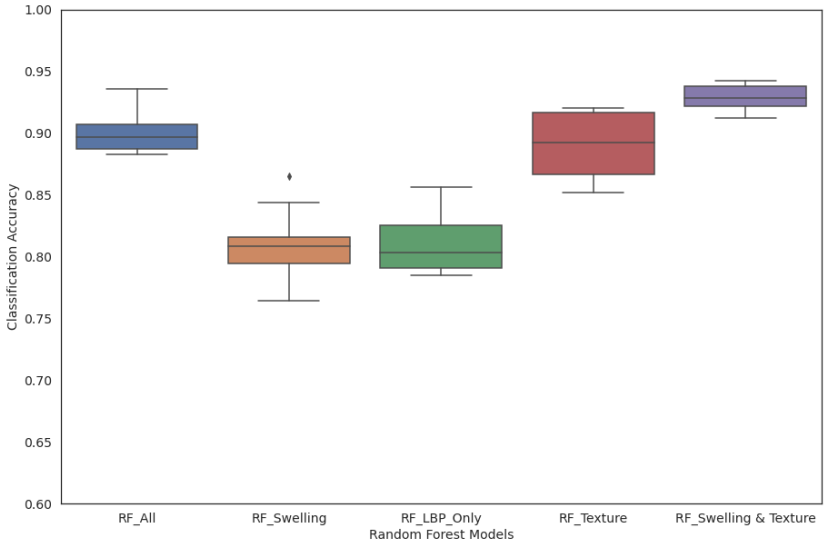

Fig. 8. Comparison of Random Forest Models based on different category groups. (a) All Features (RF_All), (b) All Swelling Features (RF_Swelling), (c) Local Binary Patterns Only (RF_LBP_Only), (d) All Texture Features (RF_Texture), (e) All Swelling & Texture Features (RF_Swelling & Texture).

swelling indicators (RF_Swelling) and local binary patterns alone (RF_LBP_Only) predicted poorly against the base model with 36 features (i.e. RF_All) due to missing information. When all texture features at the radial bone were used (RF_Texture), it achieved similar accuracy as the base model. Despite the texture features captured most of the information needed for the classification task, this resulted in a significant variation in accuracy during cross-validation subjecting to the random choice of features selected for each tree in the random forest model.

In the final model (RF_Swelling & Texture), other indicator groups (e.g. swelling) were used to provide supplemental information with an aim to boost accuracy. The better performance against the base model indicated that relevant supplementary features could enhance performance despite its weak predictability on its own. It however also implied that the patient's age, gender and indication of osteoporosis were not the attributing features for this classification task and instead hampered the random selection of the most important features used by the random forest model (i.e. lower accuracy observed in the base model.)

4 Conclusion

This work has presented the complexity and multivariate nonlinearity for treatment procedure classification. Random Forest model was shown as the model

of choice for this nonlinear tasks where weak classifications from decision trees were ensembled with an average classification accuracy of 90.0%.

Based on the analysis of the input data and model outcome, this work has also found similar feature importance as [21] where texture indicators at the radial bones were crucial to the classification of the input images. Local Binary Profile, together with the intensity changes across landmark points, essentially provided information on the presence of edges (i.e. signs of fracture). Other features, e.g. swelling, would aid prediction but did not classify well on its own.

From these findings, it is expected that a convolution neural network (CNN) models for classification will yield similar, if not better, performance given the mechanism of feature extraction across a CNN architecture. This will form the basis for investigation in future work.

References

1. Arora, R., Gabl, M., Gschwentner, M., Deml, C., Krappinger, D., Lutz, M.: A Comparative Study of Clinical and Radiologic Outcomes of Unstable Colles Type Distal Radius Fractures in Patients Older Than 70 Years: Nonoperative Treatment Versus Volar Locking Plating. Journal of Orthopaedic Trauma **23**(4), 237–242 (Apr 2009). https://doi.org/10.1097/BOT.0b013e31819b24e9
2. Barai, A., Lambie, B., Cosgrave, C., Baxter, J.: Management of distal radius fractures in the emergency department: A long-term functional outcome measure study with the Disabilities of Arm, Shoulder and Hand (DASH) scores. Emergency Medicine Australasia **30**(4), 530–537 (2018). https://doi.org/10.1111/1742-6723.12946
3. Bartl, C., Stengel, D., Bruckner, T., Rossion, I., Luntz, S., Seiler, C., Gebhard, F.: Open reduction and internal fixation versus casting for highly comminuted and intra-articular fractures of the distal radius (ORCHID): protocol for a randomized clinical multi-center trial. Trials **12**(1), 84 (Mar 2011). https://doi.org/10.1186/1745-6215-12-84
4. Bidgood, W.D., Horii, S.C.: Introduction to the ACR-NEMA DICOM standard. Radiographics: A Review Publication of the Radiological Society of North America, Inc **12**(2), 345–355 (Mar 1992). https://doi.org/10.1148/radiographics.12.2.1561424
5. Bishop, C.: Pattern Recognition and Machine Learning. Information Science and Statistics, Springer-Verlag, New York (2006), https://www.springer.com/gp/book/9780387310732
6. Bloom, R.A., Laws, J.W.: Humeral cortical thickness as an index of osteoporosis in women. The British Journal of Radiology **43**(512), 522–527 (Aug 1970). https://doi.org/10.1259/0007-1285-43-512-522
7. Breiman, L.: Random Forests. Machine Learning **45**(1), 5–32 (Oct 2001). https://doi.org/10.1023/A:1010933404324
8. Colles, A.: On the fracture of the carpal extremity of the radius. The New England Journal of Medicine, Surgery and Collateral Branches of Science **3**(4), 368–372 (1814). https://doi.org/10.1056/NEJM181410010030410
9. Cooney, W.P., Dobyns, J.H., Linscheid, R.L.: Complications of Colles' fractures. The Journal of Bone and Joint Surgery. American Volume **62**(4), 613–619 (1980)

10. Criminisi, A., Shotton, J.: Decision Forests for Computer Vision and Medical Image Analysis. Springer London, London (2013)
11. Grewal, R., MacDermid, J.C., King, G.J.W., Faber, K.J.: Open Reduction Internal Fixation Versus Percutaneous Pinning With External Fixation of Distal Radius Fractures: A Prospective, Randomized Clinical Trial. Journal of Hand Surgery **36**(12), 1899–1906 (Dec 2011). https://doi.org/10.1016/j.jhsa.2011.09.015
12. Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition. Springer Series in Statistics, Springer-Verlag, New York, 2 edn. (2009). https://doi.org/10.1007/978-0-387-84858-7
13. Hsu, H., Fahrenkopf, M.P., Nallamothu, S.V.: Wrist Fracture. In: StatPearls. StatPearls Publishing, Treasure Island (FL) (2020), http://www.ncbi.nlm.nih.gov/books/NBK499972/
14. Jantzen, C., Cieslak, L.K., Barzanji, A.F., Johansen, P.B., Rasmussen, S.W., Schmidt, T.A.: Colles' fractures and osteoporosis–A new role for the Emergency Department. Injury **47**(4), 930–933 (Apr 2016). https://doi.org/10.1016/j.injury.2015.11.029
15. Kapoor, H., Agarwal, A., Dhaon, B.K.: Displaced intra-articular fractures of distal radius: a comparative evaluation of results following closed reduction, external fixation and open reduction with internal fixation. Injury **31**(2), 75–79 (Mar 2000). https://doi.org/10.1016/S0020-1383(99)00207-7
16. Laseter, G.F., Carter, P.R.: Management of Distal Radius Fractures. Journal of Hand Therapy **9**(2), 114–128 (Apr 1996). https://doi.org/10.1016/S0894-1130(96)80070-6
17. Marshall, R.J.: The use of classification and regression trees in clinical epidemiology. Journal of Clinical Epidemiology **54**(6), 603–609 (Jun 2001). https://doi.org/10.1016/S0895-4356(00)00344-9
18. Muller, R., Möckel, M.: Logistic regression and cart in the analysis of multimarker studies. Clinica Chimica Acta **394**(1), 1 – 6 (2008). https://doi.org/10.1016/j.cca.2008.04.007
19. Ojala, T., Pietikäinen, M., Harwood, D.: A comparative study of texture measures with classification based on featured distributions. Pattern Recognition **29**(1), 51–59 (Jan 1996). https://doi.org/10.1016/0031-3203(95)00067-4
20. Podgorelec, V., Kokol, P., Stiglic, B., Rozman, I.: Decision Trees: An Overview and Their Use in Medicine. Journal of Medical Systems **26**(5), 445–463 (Oct 2002). https://doi.org/10.1023/A:1016409317640
21. Reyes-Aldasoro, C.C., Ngan, K.H., Ananda, A., Garcez, A.d., Appelboam, A., Knapp, K.M.: Geometric Semi-automatic Analysis of Colles' Fractures. medRxiv p. 2020.02.18.20024562 (Feb 2020), https://www.medrxiv.org/content/10.1101/2020.02.18.20024562v1
22. Shaikhina, T., Lowe, D., Daga, S., Briggs, D., Higgins, R., Khovanova, N.: Decision tree and random forest models for outcome prediction in antibody incompatible kidney transplantation. Biomedical Signal Processing and Control **52**, 456–462 (Jul 2019). https://doi.org/10.1016/j.bspc.2017.01.012
23. Simic, P.M., Weiland, A.J.: Fractures of the Distal Aspect of the Radius: Changes in Treatment Over the Past Two Decades. JBJS **85**(3), 552–564 (Mar 2003)
24. Webber, T., Patel, S.P., Pensak, M., Fajolu, O., Rozental, T.D., Wolf, J.M.: Correlation Between Distal Radial Cortical Thickness and Bone Mineral Density. The Journal of Hand Surgery **40**(3), 493–499 (Mar 2015). https://doi.org/10.1016/j.jhsa.2014.12.015